Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Statistical Planning and Inference 129 (2005) 207–227
journal of statistical planning and inference
www.elsevier.com/locate/jspi

# A new variance estimator for a two-phase restratified sample with PPS sampling at both phases

Suojin Wang[a,*], Alan H. Dorfman[b], Lawrence R. Ernst[b]

[a] *Department of Statistics, Texas A&M University, College Station, Texas 77843, USA*
[b] *Bureau of Labor Statistics, 2 Massachusetts Avenue, NE, Washington, DC 20212, USA*

Available online 21 August 2004

**Abstract**

In this paper we consider variance estimation for population totals and ratios in complex, cross-stratified surveys in which the ultimate sampling weights are random variables, dependent on the first phase of sampling. A new hybrid variance estimator, dependent on both model-based and design-based ideas, is introduced. Theoretical and empirical justifications are given which demonstrate that the proposed method handles well the difficult aspects of this sample design.


*Keywords:* Balanced half-sample method; Jackknife method; Random weighting; Superpopulation; Two-stage sampling

## 1. Introduction

In this paper we consider variance estimation for population totals and ratios in complex, cross-stratified surveys in which the ultimate sampling weights are random variables, dependent on the first phase of sampling (cf. Ernst, 1989). Such designs are useful where, for reasons of cost or logistics, it is expedient to sample primary sample units (PSUs) within one dimension of the stratification (say, area strata) and secondary units within the second dimension of the stratification (say industrial strata), the choice of secondary units being limited to the PSUs already selected. This is the situation, for example, in the US National

* Corresponding author. Tel.: +1-979-845-3141; fax: +1-979-845-3144.
*E-mail address:* sjwang@stat.tamu.edu (S. Wang).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.06.048

Compensation Survey (NCS) used by the US Bureau of Labor Statistics to estimate wage levels for occupational groups. A description of this survey can be found "on the Web" at www.bls.gov/ncs/home.htm; in the NCS, the PSUs are either Metropolitan Statistical Areas (MSAs), groups of MSAs, or single non-metropolitan counties, the area strata are collections of (1 or more) PSUs, and the secondary units are establishments selected within industrial strata.

The particular survey design and estimation we shall describe can be thought of as two stage sampling with complications, or, following the lead of Särndal et al. (1992, Chapter 9), as a variant of two-phase sampling; see, e.g., Cochran (1977, Chapter 12), Kott and Stukel (1997), Kim et al. (2000), and Binder et al. (2000). In its use of a stratification scheme for the ultimate sample units which ignores the strata used to select PSUs, it is similar to the design discussed in Kott (1990). However, it has complications that cast doubt on the applicability of currently available variance estimation methods; in particular, at both phases, it employs nonmeasurable (cf. Särndal et al. (1992, Chapter 9)) stratified pps sampling, going beyond the scope of the papers cited. We discuss this further in Section 6.

To fix ideas we borrow terminology from the NCS and employ the following notation:

$i$ = index for areas ($i = 1, \ldots, I$),
$j$ = index for industries ($j = 1, \ldots, J$),
$\ell$ = index for PSUs within area $i$ ($\ell = 1, \ldots, L_i$),
$k$ = index for establishments within PSU ($k$=1, …, $K_{ij\ell}$ for PSU $\ell$ in area $i$ and industry $j$),
$E_{ijk\ell}$ = the employment level for establishment $k$ in PSU $\ell$, area $i$ and industry $j$,
$Y_{ijk\ell}$ = a variable of interest,
$N_j$ = number of total establishments in industry $j$,
$N = \sum_j N_j$, the overall population size.

Of interest is estimating the overall population total $T = \sum_j \sum_i \sum_\ell \sum_k Y_{ijk\ell}$ or the ratio of two such population totals, as well as the corresponding quantities for each industry. The goal in this paper is to provide proper variance estimates for their point estimators.

We now briefly describe the sample design. Let $n_j$ be the number of establishments sampled in industry $j$, assumed to be known, and $n = \sum_j n_j$, the total number of sampled establishments. Note that this is different from situations where the desired sample size per industry $n_j$ depends on the results of the first phase of sampling, as in, for example, Folsom et al. (1987).

PSUs are selected with certainty when there is one PSU per area stratum, and, otherwise are considered non-certainties. We assume only one PSU is selected from each area. Non-certainty PSUs are selected by probability proportional to employment level, within the stratum, $E_{i\cdot\cdot\ell} = \sum_{jk} E_{ijk\ell}$.

Let

$$\pi_{i\ell} = \frac{\text{employment level in the } (i,\ell)\text{th PSU}}{\text{employment level in the } i\text{th area}} = \frac{E_{i\cdot\cdot\ell}}{\sum_u E_{i\cdot\cdot u}},$$

$\ell_i$ = index for the sampled PSU in area $i$,

$\alpha_{ijk\ell_i} = \pi_{i\ell_i}^{-1} E_{ijk\ell_i}$.

Then $\ell_i$ and $\alpha_{ijk\ell_i}$ are random variables. The $\alpha_{ijk\ell_i}$ are used as size measures for selecting secondary units (establishments) within industrial stratum $j$.

When there are no certainty establishments, define

$$\pi_{ijk\ell_i} = \frac{n_j \alpha_{ijk\ell_i}}{\sum_{r,t} \alpha_{rjt\ell_r}} \in (0, 1), \tag{1}$$

When there are certainty establishments, collect the establishments that have $\pi_{ijk\ell_i}$ in (1) $\geqslant 1$, and call the collection $C_j$. Let $n_{C_j}$ be the subsample size assigned to $C_j$ proportional to its size, and $n_j^* = n_j - n_{C_j}$. Define

$$\pi_{ijk\ell_i} = \begin{cases} 1 & \text{for } (i,k) \in C_j, \\ \dfrac{n_j^* \alpha_{ijk\ell_i}}{\sum_{\overline{C}_j} \alpha_{rjt\ell_r}} & \text{for } (i,k) \in \overline{C}_j. \end{cases} \tag{2}$$

If some of $\pi_{ijk\ell_i}$ are still $\geqslant 1$, repeat the standard procedure above until all of them are $< 1$.

The sampling procedure is:

(1) use inclusion probabilities $\pi_{i\ell}$ to select one PSU (certainty or not) per area;
(2) for each industry, conditional on the sampled PSUs sample establishments using systematic *pps* sampling with inclusion probabilities $\pi_{ijk\ell_i}$.

Here we assume that at phase 1, sampling of a PSU within each area is independent of sampling in other areas and that phase 2 (given phase 1) sampling of establishments within each industry is independent of sampling in other industries, at least approximately at both phases.

Let

$$w_{ijk\ell} = \begin{cases} (\pi_{ijk\ell_i}\pi_{i\ell})^{-1} & \text{if } (i,j,k,\ell) \text{ is in sample,} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Once we have obtained the sampled data, we use

$$\widehat{T} = \sum_j \sum_i \sum_\ell \sum_k w_{ijk\ell} Y_{ijk\ell} = \sum_j \widehat{T}_j \tag{4}$$

to estimate $T$, where $\widehat{T}_j = \sum_i \sum_k w_{ijk\ell_i} Y_{ijk\ell_i}$. It can be shown that $\widehat{T}$ is unbiased for $T$. A proof of this is given below. The idea follows the work of Ernst (1989).

Note that $w_{ijk\ell}$ is a random variable. It suffices to show that $E(w_{ijk\ell}) = 1$ for all $i, j, k, \ell$. This is a key desirable property for random weighted designs. We have

$$\begin{aligned} E(w_{ijk\ell}) &= E\{E(w_{ijk\ell}|\ell_1, \ldots, \ell_I \text{ in sample})\} \\ &= E\{(\pi_{ijk\ell_i}\pi_{i\ell})^{-1} P[(i,j,k,\ell) \text{ sampled } | \ell_i = \ell, \\ &\quad \ell_1, \ldots, \ell_I \text{ in sample}] \cdot P(\ell_i = \ell | \ell_1, \ldots, \ell_I \text{ in sample})\} \\ &= E\{(\pi_{ijk\ell_i}\pi_{i\ell})^{-1} \pi_{ijk\ell_i}\pi_{i\ell}\} = 1, \end{aligned}$$

completing the proof.

One reason behind this sampling scheme is that in the special case when $Y_{ijk\ell} = E_{ijk\ell}$, the design variable, we have the desirable property that $\widehat{T} = T$ a.s. To show this, we

observe that

$$\begin{aligned}\widehat{T}_j &= \left(\sum_{(i,k)\in C_j} + \sum_{(i,k)\in \overline{C}_j}\right) w_{ijk\ell_i} E_{ijk\ell_i} \\ &= \sum_{(i,k)\in C_j} \frac{1}{\pi_{i\ell_i}} E_{ijk\ell_i} + \sum_{(i,k)\in \overline{C}_j} \frac{\pi_{i\ell_i} \sum_{(r,t)\in \overline{C}_j} \alpha_{rjt\ell_r}}{n_j^* E_{ijk\ell_i}} \frac{1}{\pi_{i\ell_i}} E_{ijk\ell_i} \\ &= \sum_{(i,k)\in C_j} \frac{1}{\pi_{i\ell_i}} E_{ijk\ell_i} + \sum_{(r,t)\in \overline{C}_j} \alpha_{rjt\ell_r} \\ &= \sum_i \frac{1}{\pi_{i\ell_i}} E_{ij\cdot\ell_i}\end{aligned}$$

a.s. since $\sum_{(r,t)\in \overline{C}_j} \alpha_{rjt\ell_r} = \sum_{(r,t)\in \overline{C}_j} E_{rjt\ell_r}/\pi_{r\ell_r}$ a.s. Therefore, $\widehat{T} = \sum_j \widehat{T}_j = \sum_i E_{i\cdot\cdot\ell_i}/\pi_{i\ell_i} = \sum_i \sum_u E_{i\cdot\cdot u} = T$ a.s., as was to be shown. However, in this case it is seen that $\widehat{T}_j$ is generally not equal to $T_j$.

Now the main question is how to estimate the variance of $\widehat{T}$. Among the difficulties associated with this particular sampling scheme, there is only one PSU per area. There are also various correlations among the terms in $\widehat{T}$. We are also interested in estimating the variance of industrywide estimators $\widehat{T}_j$. Existing methods for two-phase sampling do not appear to be readily applicable to our problem (see Section 6). One difficulty, the fact of one PSU per area stratum is frequently handled in practice by an ad hoc collapsed stratum approach. We propose instead a model-based approach for the first-phase variance component, and use a design-based approach for estimating the second-phase variance component, leading to an overall hybrid estimator.

The rest of the paper is organized as follows. Some existing methods for variance estimation are reviewed briefly in Section 2. In Section 3, a new variance estimator is proposed for $\widehat{T}$ defined in (4) together with its analytic justifications. Some results of a simulation study are given in Section 4. The methodology is extended to the problem of estimating population ratios in Section 5 with additional simulation results. Some concluding remarks are given in Section 6.

## 2. Some existing methods

We here discuss two replication methods of variance estimation. It is not intended to suggest that these are particularly appropriate to the sample design in question. Indeed work of Kott and Stukel (1997) would tend to suggest they are not. Nonetheless, because of their relative ease of application they seem to us to be worthy of consideration and comparison, and at this point in time, as we note below, no clearly better replication methods appear to be available.

### 2.1. Balanced half-sample variance estimation

One possible approach is to apply a balanced half-sample (BHS) variance estimation procedure proposed by McCarthy (1969) to this problem; see also Valliant et al. (2000, Section 10.2). We describe the BHS methodology here briefly. It should be noted that the BHS and jackknife variance estimators require more than one PSU per (area) stratum. The BHS demands two PSUs per stratum, and to implement it, we conjoined "adjacent" pairs of strata into single "variance strata", adjacent being judged by supposed (and in the simulation, actual) likeness in the relationship of $Y$ to $E$ in the areas. For the jackknife we used a "delete one stratum at a time" approach. These are perhaps over simple ad hoc procedures, but with some support from "tradition" behind them (compare Wolter (1985, Section 2.5)), and would be readily implemented, if it could be shown, for particular data, they do not give estimates that are too unreasonable.

For simplicity, let $I$ be even. Some adjustments will be required when $I$ is odd. We first divide the $I$ sampled PSUs into $H = [I/2]$ groups. Let $Q$ be the number of minimal half samples that are in full orthogonal balance. Then $H+1 \leqslant Q \leqslant H+4$ (Valliant et al. (2000, p. 331)). A minimal set of half-samples in full orthogonal balance can be obtained by properly using a Hadamard matrix. For example, in our simulation study reported later, $H = 15$, $Q = 16$, and columns 2–16 of the $16 \times 16$ Hadamard matrix are used to construct such a set of half-samples, where each row of the submatrix corresponds to a unique half sample.

For the $q$th half-sample, we define estimator $\widehat{T}^{(q)}$ for estimating the population total $T$ as follows:

$$\begin{aligned}\widehat{T}^{(q)} &= 2\sum_j \sum_i \sum_l \sum_k w_{ijk\ell}^{(q)} Y_{ijk\ell} \\ &= 2\sum_j \sum_i \sum_k w_{ijk\ell_i}^{(q)} Y_{ijk\ell_i} \\ &= \sum_j \widehat{T}_j^{(q)},\end{aligned}$$

where $w^{(q)}$ are the original weights given in (3) for the establishments in the $q$th half-sample and 0 otherwise, and $\widehat{T}_j^{(q)} = 2\sum_i \sum_k w_{ijk\ell_i}^{(q)} Y_{ijk\ell_i}$. The new weights $2w^{(q)}$ are chosen intuitively and conveniently to reflect the half-sample size relative to the full sample size in $\widehat{T}$. Using alternative weights is possible. For example, replacing $2w_{ijk\ell_i}^{(q)}$ by $w_{ijk\ell_i}^{(q)} \sum_m \sum_p w_{mjp\ell_m} / \sum_m \sum_p w_{mjp\ell_m}^{(q)}$ has been considered in our simulation study with little improvement in the variance estimation.

The BHS variance estimator is then defined as

$$v_{\mathrm{BHS}} = \frac{1-f}{Q} \sum_{q=1}^{Q} (\widehat{T}^{(q)} - \widehat{T})^2, \tag{5}$$

where $f = n/N$ and $1 - f$ is an approximate finite population correction factor in the half-sample sampling context. Using a more complicated correction factor is possible, but it would lead to little numerical difference in our setting.

### 2.2. Jackknife variance estimation

A second approach is to use the jackknife methodology first introduced by Quenouille (1949) and Tukey (1958). Note that in selecting half-samples in the BHS procedure half of the areas are chosen in each half-sample. A jackknife variance estimator can be constructed similarly. Let $\widehat{T}_{(i)}$ be the delete $i$ estimator of $T$, i.e.,

$$\widehat{T}_{(i)} = \frac{I}{I-1} \sum_{j} \sum_{m \neq i} \sum_{k} w_{mjk\ell_m} Y_{mjk\ell_m}$$

and $\widehat{T}_{\text{Jack}} = \sum_i \widehat{T}_{(i)}/I$. Then the standard jackknife variance estimator is defined as

$$v_{\text{Jack}} = (1-f)\left(1 - \frac{1}{I}\right) \sum_{i=1}^{I} (\widehat{T}_{(i)} - \widehat{T}_{\text{Jack}})^2. \tag{6}$$

As will be seen later, we have examined the performance of $v_{\text{BHS}}(\widehat{T})$ and $v_{\text{Jack}}(\widehat{T})$ in our simulation study. We have also attempted to use modified weights other than those used in (5) and (6), but the performance does not appear to be much improved. One option for the jackknife we tried, which was later also proposed by a referee, replaces $I/(I-1)$ in $\widehat{T}_{(i)}$ by $E/(E - E_{i\cdots})$, where $E = \sum_i E_{i\cdots}$.

Following the ideas used in (5) and (6), it is quite straightforward to define the BHS and jackknife variance estimators for estimated industrywide totals. For the $j$th industry total $T_j$, the variance of its point estimator $\widehat{T}_j$ given in (4) may be estimated by dropping the summation over the industries ($j$) in (5) and (6). The resultant BHS and jackknife variance estimators are

$$v_{\text{BHS},j} = \frac{1-f_j}{Q} \sum_{q=1}^{Q} (\widehat{T}_j^{(q)} - \widehat{T}_j)^2 \tag{7}$$

and

$$v_{\text{Jack},j} = (1-f_j)\left(1 - \frac{1}{I}\right) \sum_{i=1}^{I} (\widehat{T}_{j(i)} - \widehat{T}_{\text{Jack},j})^2, \tag{8}$$

where each subscript $j$ indicates estimation specifically for the $j$th industry and $f_j = n_j/N_j$ is used to approximate the finite population correction.

## 3. A new variance estimator

We now propose a new hybrid method for the variance estimation problem that makes use of both design-based and model-based strategies. First, we have

$$\begin{aligned} V &= \mathrm{var}(\widehat{T}) \\ &= E\{\mathrm{var}(\widehat{T}\,|\,\text{phase 1})\} + \mathrm{var}\{E(\widehat{T}\,|\,\text{phase 1})\} \\ &= \sum_j E\{\mathrm{var}(\widehat{T}_j\,|\,\text{phase 1})\} + \sum_i \mathrm{var}\left(\sum_{\ell=1}^{L_i} \frac{\delta_{i\ell} Y_{i\cdot\cdot\ell}}{\pi_{i\ell}}\right) \\ &= A + B, \end{aligned} \tag{9}$$

where $\delta_{i\ell} = 1$ if $\ell = \ell_i$ and 0 otherwise. In the third equation above, the first term is due to the zero correlation assumption of phase 2 sampling (given phase 1) across the industries, while the second term uses the zero correlation assumption of phase 1 sampling across the areas. Therefore, we have partitioned $V$ into two variance components: $A$ the second-phase variance and $B$ the first-phase variance.

Assume that the population listing of establishments can be regarded as approximately random within each industry. Recall that the second-phase sampled units are obtained by systematic pps sampling given the sampled PSUs from first phase of sampling. Then for each $j$, $A_j = E\{\mathrm{var}(\widehat{T}_j\,|\,\text{phase 1})\}$ may be estimated by several methods given in Wolter (1985), such as the Yates and Grundy-type estimator (described as $v_9$ in Wolter (1985, p. 287)), or an alternative ($v_{10}$ in Wolter (1985, p. 287)) by treating the second-phase sampling as if it were with replacement. Therefore, the first term in (9) is estimable. In this work, we choose to use the alternative estimator (with a finite population correction term $1 - f_j$ for each $j$) since it is simple to implement and it appears to fit well our framework. It tends to be conservative. In many applications including the NCS, the $f_j$'s are negligibly small. The resulting estimator for $A$ is

$$\widehat{A} = \sum_j \frac{(1-f_j)n_j}{n_j - 1} \sum_{(i,k)\ \text{in sample}} (w_{ijk\ell_i} Y_{ijk\ell_i} - \widehat{\overline{Y}}_{\cdot j\cdot\cdot})^2, \tag{10}$$

where $\widehat{\overline{Y}}_{\cdot j\cdot\cdot}$ = the sample mean of $w_{ijk\ell_i} Y_{ijk\ell_i}$ for each $j$.

The remaining question is how to estimate the second term $B$ in (9) which measures the variation of the first-phase sampling. Let $Z_i = Y_{i\cdot\cdot\ell_i}/\pi_{i\ell_i}$. There is only one $Y_{i\cdot\cdot\ell_i}$ for each area, which makes it difficult to estimate $\mathrm{var}(Z_i)$ using only the observed information from the $i$th area. The standard Horvitz–Thompson based estimators have the same problem as the replication estimators in this case. Furthermore, the $Z_i$ is unobservable since $Y_{i\cdot\cdot\ell_i}$ is the triple sum over $j, k, \ell$, not just the sampled total.

To overcome this problem we propose a hybrid approach that makes use of a stipulated model for the population which is embedded in a superpopulation. Assume that $(Y_{ijk\ell}, E_{ijk\ell})$ follow a working superpopulation model:

$$\begin{aligned} Y_{ijk\ell} &= \mu_{ijk\ell} + \varepsilon_{ijk\ell} \\ &= g_j(E_{ijk\ell}) + \varepsilon_{ijk\ell}, \end{aligned} \tag{11a}$$

where $\mu_{ijk\ell} = g_j(E_{ijk\ell})$ are the mean functions in terms of $E_{ijk\ell}$, $\varepsilon_{ijk\ell}$ are independent random errors with mean 0 and variance $\sigma_{ij}^2$. For simplicity, we assume that $\sigma_{ij}^2 = \sigma_j^2$ and $g_j$ are polynomial functions, say quadratic:

$$g_j(E) = \beta_{j0} + \beta_{j1}E + \beta_{j2}E^2. \tag{11b}$$

The design variance of $W = \sum_{i\ell} \delta_{i\ell} Y_{i\cdot\cdot\ell}/\pi_{i\ell}$ in (9) is

$$\begin{aligned} B &= \sum_{i=1}^{I} \operatorname{var}\left( \sum_{\ell=1}^{L_i} \frac{\delta_{i\ell} Y_{i\cdot\cdot\ell}}{\pi_{i\ell}} \right) \\ &= \sum_{i=1}^{I} \sum_{\ell=1}^{L_i} \frac{Y_{i\cdot\cdot\ell}^2}{\pi_{i\ell}} - \sum_{i=1}^{I} Y_{i\cdots}^2. \end{aligned} \tag{12}$$

With the setup of model (11), we can estimate both terms in (12). Here, we continue to use a dot to denote the sum over a particular index as before. Then from (11) we have

$$Y_{i\cdots} = \mu_{i\cdots} + \varepsilon_{i\cdots} \approx \mu_{i\cdots} \tag{13}$$

and

$$Y_{i\cdot\cdot\ell}^2 = \mu_{i\cdot\cdot\ell}^2 + 2\mu_{i\cdot\cdot\ell}\varepsilon_{i\cdot\cdot\ell} + \varepsilon_{i\cdot\cdot\ell}^2,$$

so that

$$\sum_{i=1}^{I} \sum_{\ell=1}^{L_i} \frac{Y_{i\cdot\cdot\ell}^2}{\pi_{i\ell}} \approx \sum_{i=1}^{I} \sum_{\ell=1}^{L_i} \frac{1}{\pi_{i\ell}} \left( \mu_{i\cdot\cdot\ell}^2 + \sum_{j=1}^{J} K_{ij\ell}\sigma_j^2 \right). \tag{14}$$

The approximations in (13) and (14) are due to the Central Limit Theorem which indicates that $\sum_{jk\ell} \varepsilon_{ijk\ell}$ and $\sum_{i\ell} \mu_{i\cdot\cdot\ell}\varepsilon_{i\cdot\cdot\ell}/\pi_{i\ell}$ are small relative to their dominating terms, respectively, as $K_{ij\ell}$ and/or $L_i$ increase. Furthermore, in (14) we have used the approximation $\varepsilon_{i\cdot\cdot\ell}^2 \approx \sum_j K_{ij\ell}\sigma_j^2$, the latter being the expected value of the former. This term is typically smaller relative to the dominating term, also due to the CLT, but we keep this term to avoid a negative bias in the resulting estimator.

Note that while the independence of errors $\varepsilon_{ijk\ell}$ in (11) leads to approximations (13) and (14), it requires only more relaxed covariance structure among the errors for the approximations to be still valid. Moreover, the assumption of constant error variance within each industry is more for convenience than for necessity. These points are demonstrated in the simulation study in Section 4.

Now $B$ in (12) may be approximated by

$$B^* = \sum_{i=1}^{I} \sum_{\ell=1}^{L_i} \frac{1}{\pi_{i\ell}} \left( \mu_{i\cdot\cdot\ell}^2 + \sum_{j=1}^{J} K_{ij\ell}\sigma_j^2 \right) - \sum_{i=1}^{I} \mu_{i\cdots}^2. \tag{15}$$

The $\mu_{i\cdot\cdot\ell}$, $\mu_{i\cdots}$ and $\sigma_j^2$ are generally unknown, but they can be estimated. Since the second-phase inclusion probabilities depend on the outcome variable only through $E$-variables

which are in fact conditioned in the model, the second-phase design is ignored in getting a model-based variance formula in (15). First we fit regression model (11) properly for each $j$ by the least-squares method. Then we obtain the sums of the fitted values $\widehat{\mu}_{i\cdot\cdot\ell} = \sum_{jk} \widehat{\mu}_{ijk\ell}$, $\widehat{\mu}_{i\cdots} = \sum_{\ell} \widehat{\mu}_{i\cdot\cdot\ell}$ and variance estimates $\widehat{\sigma}_j^2$. Hence, via (15), $B$ may be estimated by

$$\widehat{B} = \sum_{i=1}^{I} \sum_{\ell=1}^{L_i} \frac{1}{\pi_{i\ell}} \left( \widehat{\mu}_{i\cdot\cdot\ell}^2 + \sum_{j=1}^{J} K_{ij\ell} \widehat{\sigma}_j^2 \right) - \sum_{i=1}^{I} \widehat{\mu}_{i\cdots}^2. \tag{16}$$

Under the correct model (11), this hybrid plug-in estimator is asymptotically unbiased provided that both $n$ and $N$ become large. Moreover, even if the fitted model (11) does not provide much useful clue about the relationship between $Y$ and $E$, $\widehat{B}$ seems to continue to do well, as is suggested by our simulation results. In addition, $\widehat{B}$ is always $\geqslant 0$ since for any constants $c_{i\cdot\cdot\ell}$, $\sum_{\ell=1}^{L_i} c_{i\cdot\cdot\ell}^2/\pi_{i\ell} - c_{i\cdots}^2 = \mathrm{var}\{\sum_{\ell} \delta_{i\ell}\, c_{i\cdot\cdot\ell}/\pi_{i\ell}\} \geqslant 0$. Finally, using (16) and (10) we obtain a new estimator for $V = \mathrm{var}(\widehat{T})$:

$$v_{\mathrm{new}} = \widehat{A} + \widehat{B}. \tag{17}$$

The issue of having certainty PSUs and/or establishments is easily handled in this approach. If there is a certainty PSU in any area $i$, the variation contributed from this area to the first-phase variance $B$ is zero. Hence, such areas should be omitted from the summation over $i$ in $B$, $\widehat{B}$ and $\widehat{B}_j$ in (12), (16) and (19) below. On the other hand, if there are certainty establishments in any given industry $j$, naturally these establishments need to be excluded in computing the second-phase variance estimate $\widehat{A}_j$ in (19) and (10). Both terms in (17) are generally important numerically. For example, as we will see in the next section, the mean square root of "expected" $\widehat{B}$ is about 30% of that of $\widehat{A}$ in the first example.

The partition of $V = A + B$ is valid even though the second-phase units are not sampled independently across the selected PSUs. Since $v_{\mathrm{new}}$ is consistent for $V$ in this case, it takes into account naturally the possible correlations of the sampled data within and across the industries. Ignoring such correlations may result in simpler but perhaps significantly biased estimators. One such possibility is to use the following simple estimator for $V$:

$$\widetilde{v} = \sum_{j=1}^{J} v_{\mathrm{new},\,j}, \tag{18}$$

where $v_{\mathrm{new},\,j}$ is the new variance estimator given in (19) below, which is defined in the same fashion as $v_{\mathrm{new}}$ but for industrywide variance $V_j = \mathrm{var}(\widehat{T}_j)$. Some empirical evidence of the performance of all these estimators is provided in the simulation experiments described in Section 4.

We now discuss how to estimate $V_j$. This can be readily done by using (10) and (16) for each industry $j$. Specifically, define

$$v_{\mathrm{new},\,j} = \widehat{A}_j + \widehat{B}_j, \tag{19}$$

where

$$\widehat{A}_j = \frac{(1-f_j)n_j}{n_j-1} \sum_{(i,k)\ \text{in sample}} (w_{ijk\ell_i} Y_{ijk\ell_i} - \widehat{\overline{Y}}_{\cdot j\cdot\cdot})^2$$

and

$$\widehat{B}_j = \sum_{i=1}^{I} \sum_{\ell=1}^{L_i} \frac{1}{\pi_{i\ell}} (\widehat{\mu}_{ij\cdot\ell}^2 + K_{ij\ell}\widehat{\sigma}_j^2) - \sum_{i=1}^{I} \widehat{\mu}_{ij\cdot\cdot}^2,$$

with $\widehat{\mu}_{ij\cdot\ell}$ and $\widehat{\mu}_{ij\cdot\cdot}^2$ being the sums of the fitted values $\widehat{\mu}_{ijk\ell}$ over the indices with a dot, respectively. In the simulation study, we have also examined the performance of $v_{\text{new},j}$.

Recall that the main purpose of this work is to find accurate variance estimators of $V$ and $V_j$. However, we are also interested in constructing confidence intervals. This can be done conveniently by using the standard $t$-type confidence intervals once an estimator for the variance of $\widehat{T}$ is obtained. Empirical coverage probabilities of the confidence intervals using several variance estimators, $v_{\text{BHS}}$, $v_{\text{Jack}}$, $v_{\text{new}}$ and $\widetilde{v}$, have been investigated and are reported in the next section.

## 4. A simulation study

In this section, we describe a simulation study to compare the performance of several variance estimators discussed in the earlier sections. We start with the steps for generating an artificial complex cross-stratified population in keeping with the set up in Section 1 but with a smaller scale then what is in the NCS. All the pseudo random variables below are generated independently of each other.

*Steps to generate a finite population:*

(1) Let $I = 30$, $J = 6$, $R = 5$;
(2) Generate $L_i \sim \text{Unif}[2, 6]$ (end points inclusive), $K_{ij\ell} \sim \text{Unif}[5, 30]$;
(3) Generate $E_{ijk\ell} \sim \text{Gamma}(10, 0.2)$, where $a = 10$ is the shape parameter and $b = 0.2$ is the scale parameter, so that the mean is $a/b = 50$ and the variance is $a/b^2 = 250$;
(4) Generate $\alpha_j \sim N(15, 0.8^2)$, $\beta_j \sim N(0.05, 0.01^2)$, $\gamma_j \sim N(0.05, 0.01^2)$;
(5) Generate $\Delta_r \sim N(0, 0.8^2)$ for $r = 1, \ldots, R$, $\tau_\ell \sim N(0, 1)$, $\xi_{ijk\ell} \sim \text{Gamma}(5, 0.2)$;
(6) Compute $u_{ijk\ell} = c_j \xi_{ijk\ell} - 25$, where $(c_1, c_2, c_3, c_4, c_5, c_6) = (2, 1.5, 1, 1, .5, .5)$,

$$\varepsilon_{ijk\ell} = E_{ijk\ell}^{1/2} u_{ijk\ell}$$

and

$$Y_{ijk\ell} = \alpha_j + \beta_j E_{ijk\ell} + \gamma_j E_{ijk\ell}^2 + E_{ijk\ell}^{1/2} (\Delta_{[1+(i-1)R/I]} + \tau_\ell) + \varepsilon_{ijk\ell},$$

with $[1 + (i-1)R/I]$ being the largest integer not exceeding $1 + (i-1)R/I$ to reflect some correlations among neighboring areas. If $Y_{ijk\ell} < 5$, define $Y_{ijk\ell} = 5$;
(7) Collect all pairs $(Y_{ijk\ell}, E_{ijk\ell})$ to compose a finite population.

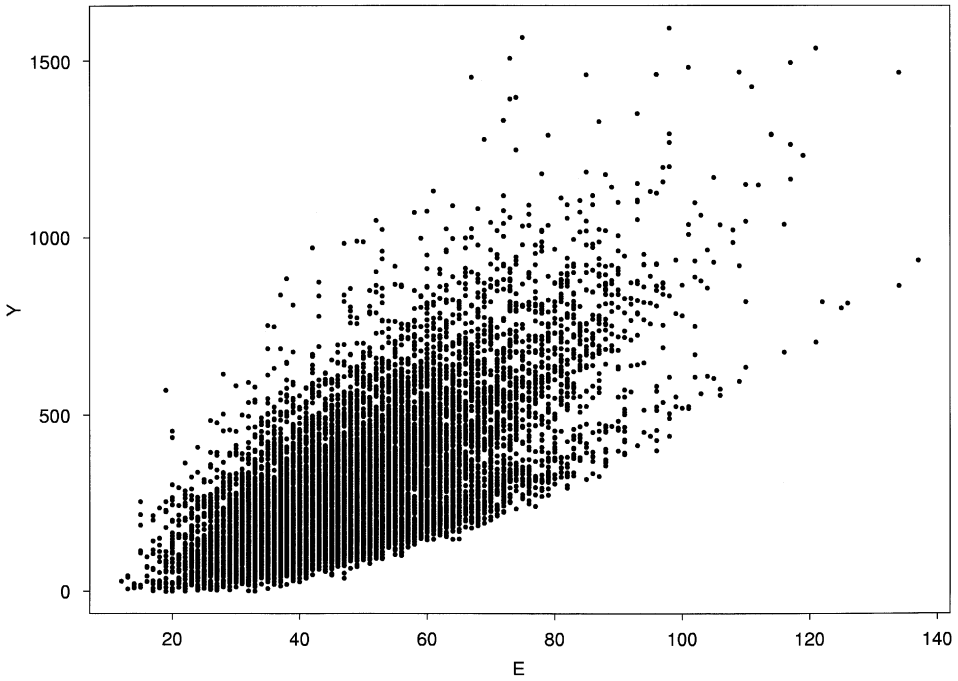


Fig. 1. Scatter plot of the first population.

A population was generated using the steps above with the overall population mean of 318.2. It reflects the possibility of various correlations of $Y$'s among industries, within areas, among neighboring areas and within PSUs. The error terms do not have constant variance and their correlations also exist among neighboring areas and PSUs. The resulting population size is $N = 10508$ with subpopulation sizes $N_1 = 1636$, $N_2 = 1724$, $N_3 = 1887$, $N_4$=1815, $N_5$=1722, $N_6$=1724. The population total is $T$=3 343 531 and the subpopulation totals are $T_1 = 788\,754$, $T_2 = 746\,859$, $T_3 = 604\,749$, $T_4 = 618\,425$, $T_5 = 269\,252$ and $T_6 = 315\,493$. The scatter plot of the population is given in Fig. 1.

After the population was obtained, 1000 samples were drawn from the population independently. In each run, a sample of size $n = 120$ was generated using the sampling scheme described in Section 1, with subsample size $n_j = 20$ for each $j$ ($j$th industry). For each sampled data set, we computed several variance estimators for the estimated population total $\widehat{T}$: $v_{\mathrm{BHS}}$, $v_{\mathrm{Jack}}$, $v_{\mathrm{new}}$, $v^*$ and $\widetilde{v}$, where

$$v^* = \frac{(1-f)n}{n-1} \sum_{(i,k,j)\ \text{in sample}} (w_{ijk\ell_i} Y_{ijk\ell_i} - \widehat{\overline{T}})^2$$

is included here for comparison also, and $\widehat{\overline{T}}$ is the overall sample mean of $w_{ijk\ell_i} Y_{ijk\ell_i}$. The "true" variance in Table 1 is the empirical variance of the 1000 realized $\widehat{T}$. Note that with

Table 1
Comparisons of variance estimators in estimation of the total in the first population

| Population | "true" $V$ | $v_{\text{BHS}}$ | $v_{\text{Jack}}$ | $v_{\text{new}}$ | $v^*$ | $\widetilde{v}$ |
|---|---|---|---|---|---|---|
| Overall | 102 277 | 327 643 | 305 572 | 106 575 | 148 130 | 140 856 |
| $j=1$ | 83 050 | 149 858 | 140 716 | 82 446 | 56 706 | N/A |
| $j=2$ | 67 388 | 134 178 | 129 434 | 68 068 | 46 763 | N/A |
| $j=3$ | 53 011 | 103 995 | 100 124 | 56 412 | 39 726 | N/A |
| $j=4$ | 57 511 | 108 794 | 105 163 | 58 897 | 37 592 | N/A |
| $j=5$ | 28 240 | 49 787 | 47 609 | 28 911 | 22 876 | N/A |
| $j=6$ | 29 979 | 55 310 | 54 979 | 30 377 | 21 308 | N/A |

The entries are the square root of the "true" variances and the averages of the estimators over 1000 runs. The "true" variances are based on 1000 realized point estimates.

1000 runs the "true" variance contains some simulation error, but it is an accurate enough approximation for the purpose of our comparisons.

Table 1 gives the averages of all these variance estimators over the 1000 runs, in the square root readings. It is seen that $v_{\text{BHS}}$ and $v_{\text{Jack}}$ are both very positively biased, with $v_{\text{BHS}}$ being even worse than $v_{\text{Jack}}$. On the other hand, $v_{\text{new}}$ appears to estimate $V$ very well, much better than all other estimators considered. The other two estimators $\widetilde{v}$ and $v^*$ both overestimate $V$ as well, although not as badly as $v_{\text{BHS}}$ and $v_{\text{Jack}}$. However, it is interesting to observe that at the industry level, $v_j^*$ (which is now reduced to $\widehat{A}_j$) seriously underestimates $V_j$.

As we mentioned in Section 2, we have tried different weights when constructing $v_{\text{BHS}}$ and $v_{\text{Jack}}$, but have not noticed any significant improvements in their performance. Note also that the empirical bias of $\widehat{T}$ and $\widehat{T}_j$ are negligible, confirming their unbiasedness. For example, the empirical bias of $\widehat{T}$ is 4158 with standard error 3234. We have also obtained the square root of estimated $E(\widehat{A})$ and $E(\widehat{B})$ to be 97 083 and 28 983, respectively.

It seems generally reasonable to assume model (11) with non-zero $\beta_{j1}$ and/or $\beta_{j2}$ for the relationship between $Y$ and $E$, as is the case for the first artificial population. On the other hand, we may want to check what might happen if $Y$ and $E$ are not related at all or the relationship between $Y$ and $E$ does not follow our working model (11) when the sampling scheme and all the variance estimation procedures stay the same. For this reason, we generated two other artificial populations in the same fashion as for the first population except for $Y_{ijk\ell}$. For the second population,

$$Y_{ijk\ell} = 298 + 6u_{ijk\ell}$$

was used. The resulting population size is $N = 10\,562$ with subpopulation sizes $N_1 = 1813$, $N_2$=1718, $N_3$=1747, $N_4$=1692, $N_5$=1854, $N_6$=1738. The population total is $T$=3 278 323 and the subpopulation totals are $T_1 = 812\,136$, $T_2 = 637\,305$, $T_3 = 526\,510$, $T_4 = 501\,974$, $T_5 = 413\,792$ and $T_6 = 386\,607$.

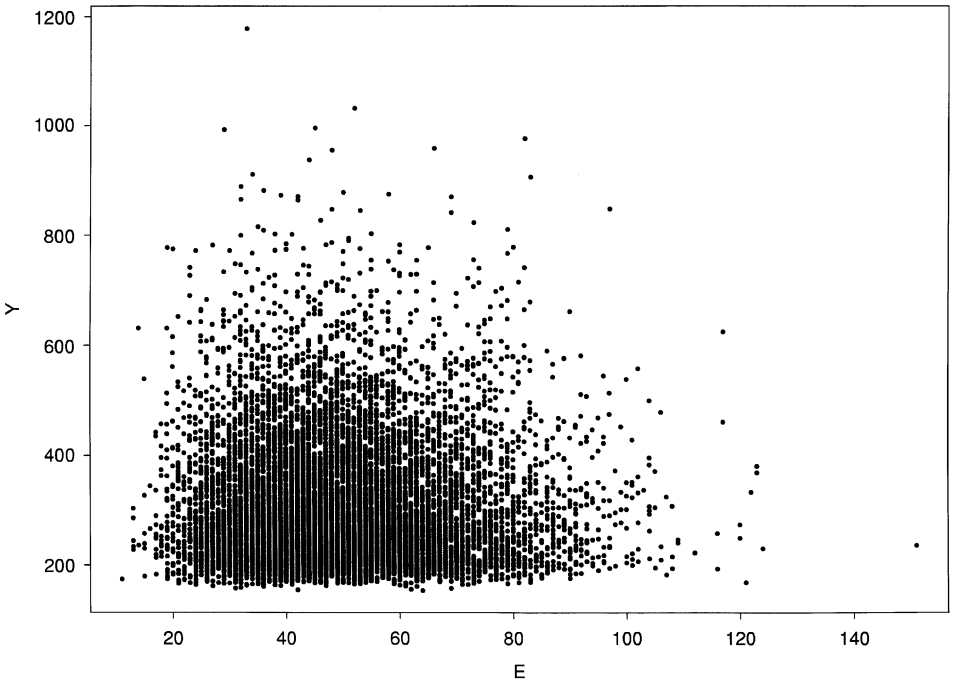


Fig. 2. Scatter plot of the second population.

For the third population,

$$Y_{ijk\ell} = 1000 - \gamma_j\{(E_{ijk\ell} - 40)/1.5\}^3 + \varepsilon_{ijk\ell}$$

was used. The resulting population size is $N = 11\,138$ with subpopulation sizes $N_1 = 1838$, $N_2 = 1878$, $N_3 = 1934$, $N_4 = 1812$, $N_5 = 1805$, $N_6 = 1871$. The population total is $T = 10\,082\,432$ and the subpopulation totals are $T_1 = 1\,952\,967$, $T_2 = 1\,812\,370$, $T_3 = 1\,786\,224$, $T_4 = 1\,575\,376$, $T_5 = 1\,441\,977$ and $T_6 = 1\,513\,519$. The scatter plots of these two populations are shown in Figs. 2 and 3. The three populations represent three quite different scenarios.

The results for the second and third populations corresponding to those in Table 1 are shown in Tables 2 and 3, respectively. It is reassuring to observe that in these different situations, similar performance of the variance estimators is shown: the new estimator $v_{\text{new}}$ again works well, while other methods often provide biased estimates for the variances of the estimated overall population total and those of the estimated subpopulation totals, although they appear to be less severe than in the case of the first population.

One might suspect that severe over estimation in the BHS and jackknife procedures might be caused by unequal inclusion probabilities $\pi_{i\ell}$ and $\pi_{ijk\ell_i}$ used in our two-phase sample design. To investigate this possibility, we generated a fourth artificial population using the same setup as that for the first population except that $L_i = 5$, $K_{ijk\ell} = 15$ and $E_{ijk\ell} = 50$ were used for all $(i, j, k, \ell)$. In this case, $\pi_{i\ell}$ and $\pi_{ijk\ell_i}$ are constants for all PSUs and establishments. Using 1000 runs, we obtained the square root of the "true" variance $V$ of

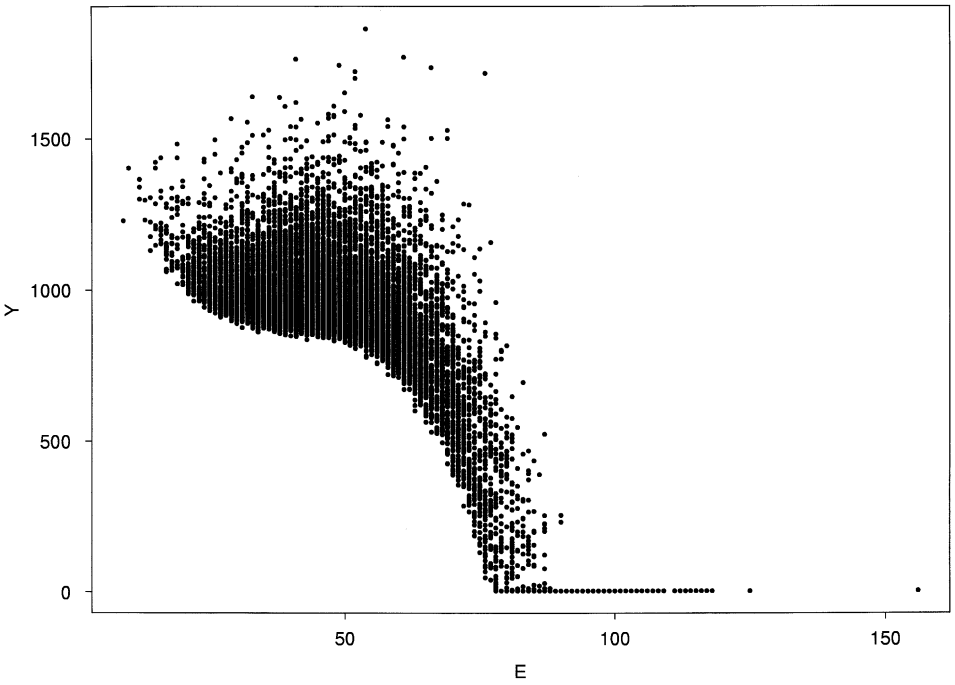


Fig. 3. Scatter plot of the third population.

Table 2
Comparisons of variance estimators in estimation of the total in the second population

| Population | "true" $V$ | $v_{\text{BHS}}$ | $v_{\text{Jack}}$ | $v_{\text{new}}$ | $v^*$ | $\widetilde{v}$ |
|---|---|---|---|---|---|---|
| Overall | 132 683 | 333 797 | 306 415 | 138 134 | 162 570 | 162 370 |
| $j=1$ | 99 534 | 159 428 | 152 011 | 98 536 | 84 872 | N/A |
| $j=2$ | 74 124 | 126 833 | 119 418 | 77 044 | 61 740 | N/A |
| $j=3$ | 64 089 | 105 331 | 98 714 | 63 525 | 50 273 | N/A |
| $j=4$ | 55 054 | 100 038 | 93 365 | 55 860 | 45 727 | N/A |
| $j=5$ | 44 067 | 76 321 | 71 672 | 43 956 | 35 266 | N/A |
| $j=6$ | 38 599 | 68 875 | 67 123 | 40 387 | 31 473 | N/A |

The entries are the square root of the "true" variances and the averages of the estimators over 1000 runs. The "true" variances are based on 1000 realized point estimates.

$\widehat{T}$ and the estimated mean of its estimates $v_{\text{BHS}}$, $v_{\text{Jack}}$, $v_{\text{new}}$, $v^*$ and $\widetilde{v}$ as 121 387, 316 638, 271 636, 120 236, 190 231 and 120 236, respectively. Here, the fitted regression values in each industry required in $v_{\text{new}}$ are reduced to be the industrywide sample mean since no regression can be done with constant $E_{ijk\ell}$. The numerical results above are similar to those in Tables 1–3, suggesting that the main cause for over estimation in $v_{\text{BHS}}$, $v_{\text{Jack}}$ and $\widetilde{v}$ is due to something other than unequal inclusion probabilities. The main difficulty with them is

Table 3
Comparisons of variance estimators in estimation of the total in the third population

| Population | "true" $V$ | $v_{\text{BHS}}$ | $v_{\text{Jack}}$ | $v_{\text{new}}$ | $v^*$ | $\widetilde{v}$ |
|---|---|---|---|---|---|---|
| Overall | 519 544 | 995 380 | 911 499 | 512 425 | 515 857 | 573 998 |
| $j=1$ | 259 823 | 382 004 | 366 140 | 266 822 | 224 163 | N/A |
| $j=2$ | 266 605 | 366 809 | 345 353 | 253 805 | 218 621 | N/A |
| $j=3$ | 216 245 | 332 960 | 320 374 | 221 029 | 190 868 | N/A |
| $j=4$ | 223 135 | 316 867 | 315 309 | 235 783 | 208 498 | N/A |
| $j=5$ | 203 748 | 287 986 | 279 385 | 210 369 | 187 016 | N/A |
| $j=6$ | 202 141 | 304 866 | 291 873 | 212 509 | 190 027 | N/A |

The entries are the square root of the "true" variances and the averages of the estimators over 1000 runs. The "true" variances are based on 1000 realized point estimates.

that no theory has been found to support their use even in this highly simplified special case with the unique two-phase sample design. Note that in this special case, $v_{\text{new}} = \widetilde{v}$. This is because $\pi_{i\ell} = 1/L_i$ and thus $\widehat{B} = \sum_{i=1}^{I}\sum_{\ell=1}^{L_i}\sum_{j=1}^{J} K_{ij\ell}\widehat{\sigma}_j^2$ and $\widehat{B}_j = \sum_{i=1}^{I}\sum_{\ell=1}^{L_i} K_{ij\ell}\widehat{\sigma}_j^2$, so that $\sum_{j=1}^{J}\widehat{B}_j = \widehat{B}$. The simulation results for variance estimates of $V(\widehat{T}_j)$ also show similar comparisons observed in Tables 1–3. The details are omitted here.

In addition to the problem of finding a proper variance estimator, we would also wish to construct a confidence interval that has a correct coverage probability. Here, we use the standard $t$-type $(1-\alpha)100\%$ confidence interval for $T$ as follows:

$$\widehat{T} \pm t_{n-3,\alpha/2}\sqrt{v},$$

where $t_{n-3,\alpha/2}$ is the $t$ critical value at level $\alpha/2$ with $n-3$ degrees of freedom and $v$ is one of the variance estimates. In our simulation examples, $n=120$, so it is essentially the same if we use the normal critical value. The empirical coverage probabilities of various confidence intervals for the overall population totals of the three artificial populations are provided in Table 4. The nominal coverage probability is 0.95. The results in Table 4 are based on 1000 runs, so that the standard errors of the empirical coverage probabilities are about .0068. It is seen that the new variance estimator performs quite well for all three populations, producing confidence intervals with length and coverage very close to those obtained by using the "true" variance. On the other hand, the other four methods are too conservative except for one case ($v^*$ in population 3). This is especially true for $v_{\text{BHS}}$ and $v_{\text{Jack}}$ which lead to exceedingly wide intervals. The coverage probabilities and interval lengths for the fourth population are similar to those given in Table 4, and are thus omitted.

## 5. Variance estimation for estimating population ratios

The methodology developed in this paper can be readily extended to the problem of variance estimation for estimating population ratios. Let $X_{ijk\ell}$ be another variable of interest

Table 4
Comparisons of the empirical coverage probabilities and lengths of the confidence intervals for the overall totals of the three populations using various variance estimators

| Population | "true" $V$ | $v_{\text{BHS}}$ | $v_{\text{Jack}}$ | $v_{\text{new}}$ | $v^*$ | $\widetilde{v}$ |
|---|---|---|---|---|---|---|
| *Coverage* | | | | | | |
| First | 0.958 | 1.000 | 1.000 | 0.967 | 0.997 | 0.994 |
| Second | 0.945 | 1.000 | 1.000 | 0.957 | 0.976 | 0.984 |
| Third | 0.947 | 1.000 | 1.000 | 0.943 | 0.947 | 0.968 |
| *Lengths* | | | | | | |
| First | 405 096 | 1 282 609 | 1 202 445 | 420 574 | 584 520 | 557 076 |
| Second | 524 945 | 1 305 333 | 1 204 150 | 542 321 | 613 683 | 639 938 |
| Third | 2 057 859 | 3 889 899 | 3 582 219 | 2 020 063 | 2 019 139 | 2 266 532 |

The nominal coverage probability is 0.95. The coverage probabilities are based on 1000 runs. The "true" variances are based on 1000 realized point estimates.

for establishment $(i, j, k, \ell)$. Suppose that we are interested in estimating the ratio

$$R = \sum_{i,j,k,\ell} Y_{ijk\ell} \Big/ \sum_{i,j,k,\ell} X_{ijk\ell} = T_Y / T_X$$

and that, by (4), we use the standard estimator

$$\widehat{R} = \sum_{j}\sum_{i}\sum_{k} w_{ijk\ell_i} Y_{ijk\ell_i} \Big/ \sum_{j}\sum_{i}\sum_{k} w_{ijk\ell_i} X_{ijk\ell_i} = \widehat{T}_Y / \widehat{T}_X$$

for $R$. One special and important case in this framework is when $X_{ijk\ell} = E_{ijk\ell}$. In general, the $X$ variable is presumably related to the $E$ variable. However, as we have seen earlier, while such a relationship is usually helpful for point estimation, it is not required for our variance estimation procedure.

The estimation of the variance of $\widehat{R}$, $V_{\widehat{R}} = \text{var}(\widehat{R})$, can be done as follows. First, by the Taylor series expansion, we obtain the usual approximation to $\widehat{R}$:

$$\widehat{R} \approx R + \frac{1}{T_X} \sum_{j}\sum_{i}\sum_{k} w_{ijk\ell_i}(Y_{ijk\ell_i} - RX_{ijk\ell_i}) = \widetilde{R}. \tag{20}$$

Then $V_{\widehat{R}}$ can be approximated by

$$\text{var}(\widetilde{R}) = \frac{1}{T_X^2} \text{var}\left( \sum_{j}\sum_{i}\sum_{k} w_{ijk\ell_i} D_{ijk\ell_i} \right), \tag{21}$$

where $D_{ijk\ell} = Y_{ijk\ell} - RX_{ijk\ell}$. Comparing the sum in (21) with that in (4), we see that if we knew $R$ then the variance in (21) can be estimated in exactly the same fashion as in (17) by replacing the $Y$'s by the $D$'s. More specifically, we have

$$V_{\widehat{R}} \approx \frac{1}{T_X^2}(A_R + B_R),$$

where $A_R$ and $B_R$ are $A$ and $B$ in (9) except that the $Y$'s are replaced by the $D$'s.

Of course, in practice, $R$ and $T_X$ are generally unknown, but their consistent estimates can be employed without affecting the asymptotic properties of the variance estimator. Using the same ideas as in Section 3, our estimator for $V_R$ is defined to be

$$v_{\text{new},R} = \frac{1}{\widehat{T}_X^2}(\widehat{A}_R + \widehat{B}_R), \tag{22}$$

where $\widehat{A}_R$ and $\widehat{B}_R$ are $\widehat{A}$ in (10) and $\widehat{B}$ in (16) except that $\widehat{D}_{ijk\ell} = Y_{ijk\ell} - \widehat{R}X_{ijk\ell}$ are used in place of $Y_{ijk\ell}$.

In estimating the industrywide ratios

$$\begin{aligned} R_j &= \sum_{i,k,\ell} Y_{ijk\ell} \Big/ \sum_{i,k,\ell} X_{ijk\ell} \\ &= T_{Yj}/T_{Xj}, \end{aligned}$$

the following point estimator

$$\begin{aligned} \widehat{R}_j &= \sum_i \sum_k w_{ijk\ell_i} Y_{ijk\ell_i} \Big/ \sum_i \sum_k w_{ijk\ell_i} X_{ijk\ell_i} \\ &= \widehat{T}_{Yj}/\widehat{T}_{Xj} \end{aligned}$$

is often used. Then our estimator for $\text{var}(\widehat{R}_j)$ is

$$v_{\text{new},R_j} = \frac{1}{\widehat{T}_{Xj}^2}(\widehat{A}_{R_j} + \widehat{B}_{R_j}), \tag{23}$$

where $\widehat{A}_{R_j}$ and $\widehat{B}_{R_j}$ are $\widehat{A}_j$ and $\widehat{B}_j$ in (19) except that $\widetilde{D}_{ijk\ell} = Y_{ijk\ell} - \widehat{R}_j X_{ijk\ell}$ are used in place of $Y_{ijk\ell}$.

Simple extensions may also be made to obtain the other four variance estimators considered in the previous sections in a parallel form. They are denoted by $v_{\text{BHS},R}$, $v_{\text{Jack},R}$, $v_R^*$ and $\widetilde{v}_R$. Note that by definition $v_R^*$ and $\widetilde{v}_R$ make use of the Taylor series expansion in (20), but the replication estimators $v_{\text{BHS},R}$, $v_{\text{Jack},R}$ are constructed with deleted versions of the raw ratio estimates without relying on the Taylor series expansion, as is reported in Tables 5–8 in the next subsection. Alternatively, a version of the BHS and jackknife estimators may be obtained by using the expansion in (20) with $\widehat{D}$ in (22). While it may be more appealing to use the raw ratio estimates, empirically neither version consistently outperforms the other.

Table 5
Comparisons of variance estimators in estimation of the ratio in the first population

| Population | "true" $V$ | $v_{\mathrm{BHS}}$ | $v_{\mathrm{Jack}}$ | $v_{\mathrm{new}}$ | $v^*$ | $\widetilde{v}$ |
|---|---|---|---|---|---|---|
| Overall | 0.195 | 0.246 | 0.247 | 0.203 | 0.282 | 0.200 |
| $j=1$ | 0.659 | 0.710 | 0.723 | 0.713 | 0.691 | N/A |
| $j=2$ | 0.530 | 0.559 | 0.567 | 0.561 | 0.545 | N/A |
| $j=3$ | 0.393 | 0.422 | 0.432 | 0.431 | 0.420 | N/A |
| $j=4$ | 0.395 | 0.413 | 0.422 | 0.425 | 0.415 | N/A |
| $j=5$ | 0.262 | 0.257 | 0.264 | 0.272 | 0.266 | N/A |
| $j=6$ | 0.248 | 0.240 | 0.247 | 0.253 | 0.246 | N/A |

The entries are the square root of the "true" variances and the averages of the estimators over 1000 runs. The "true" variances are based on 1000 realized point estimates.

Table 6
Comparisons of variance estimators in estimation of the ratio in the second population

| Population | "true" $V$ | $v_{\mathrm{BHS}}$ | $v_{\mathrm{Jack}}$ | $v_{\mathrm{new}}$ | $v^*$ | $\widetilde{v}$ |
|---|---|---|---|---|---|---|
| Overall | 0.252 | 0.267 | 0.349 | 0.262 | 0.295 | 0.261 |
| $j=1$ | 0.963 | 0.933 | 0.950 | 0.957 | 0.942 | N/A |
| $j=2$ | 0.711 | 0.707 | 0.717 | 0.733 | 0.718 | N/A |
| $j=3$ | 0.579 | 0.570 | 0.572 | 0.586 | 0.577 | N/A |
| $j=4$ | 0.544 | 0.516 | 0.527 | 0.545 | 0.535 | N/A |
| $j=5$ | 0.377 | 0.364 | 0.377 | 0.386 | 0.381 | N/A |
| $j=6$ | 0.337 | 0.338 | 0.349 | 0.366 | 0.361 | N/A |

The entries are the square root of the "true" variances and the averages of the estimators over 1000 runs. The "true" variances are based on 1000 realized point estimates.

Table 7
Comparisons of variance estimators in estimation of the ratio in the third population

| Population | "true" $V$ | $v_{\mathrm{BHS}}$ | $v_{\mathrm{Jack}}$ | $v_{\mathrm{new}}$ | $v^*$ | $\widetilde{v}$ |
|---|---|---|---|---|---|---|
| Overall | 0.919 | 0.846 | 1.974 | 0.924 | 0.914 | 0.919 |
| $j=1$ | 2.380 | 2.289 | 2.342 | 2.466 | 2.429 | N/A |
| $j=2$ | 2.442 | 2.164 | 2.194 | 2.353 | 2.311 | N/A |
| $j=3$ | 1.997 | 1.865 | 1.914 | 2.055 | 2.003 | N/A |
| $j=4$ | 2.144 | 2.074 | 2.146 | 2.337 | 2.290 | N/A |
| $j=5$ | 2.010 | 1.964 | 2.017 | 2.116 | 2.076 | N/A |
| $j=6$ | 1.998 | 1.923 | 1.974 | 2.087 | 2.046 | N/A |

The entries are the square root of the "true" variances and the averages of the estimators over 1000 runs. The "true" variances are based on 1000 realized point estimates.

### 5.1. *Additional simulation results*

In this subsection, we report the results of a simulation study to compare the performance of the aforementioned five methods for estimating the variances of $\widehat{R}$ and $\widehat{R}_j$. The simulation

Table 8
Comparisons of the empirical coverage probabilities and lengths of the confidence intervals for the overall ratios of the three populations using various variance estimators

| Population | "true" $V$ | $v_{\text{BHS}}$ | $v_{\text{Jack}}$ | $v_{\text{new}}$ | $v^*$ | $\widetilde{v}$ |
|---|---|---|---|---|---|---|
| *Coverage* | | | | | | |
| First | 0.958 | 0.973 | 0.980 | 0.967 | 0.997 | 0.964 |
| Second | 0.945 | 0.951 | 0.983 | 0.957 | 0.976 | 0.955 |
| Third | 0.943 | 0.904 | 0.999 | 0.943 | 0.945 | 0.942 |
| *Lengths* | | | | | | |
| First | 0.772 | 0.955 | 0.960 | 0.802 | 1.114 | 0.789 |
| Second | 0.998 | 1.034 | 1.341 | 1.030 | 1.159 | 1.026 |
| Third | 3.640 | 3.278 | 7.662 | 3.643 | 3.601 | 3.622 |

The nominal coverage probability is 0.95. The coverage probabilities are based on 1000 runs. The "true" variances are based on 1000 realized point estimates.

framework followed that in Section 4. We used the same three artificial populations and the same 1000 samples from each population. We let $X = E$ for the second variable of interest.

Tables 5–7 give the results of the comparisons for each of the three populations, respectively. These tables correspond to Tables 1–3 with ratios instead of totals. The true overall ratios are 6.376, 6.226 and 18.173, respectively. Formulas (21) and (22) were used for the new method in these tables, although $T_X$ and $T_{Xj}$ are available when $X = E$. However, we have used $T_X$ and $T_{Xj}$ as well and obtained identical results to the decimals in the tables for estimating $R$ and nearly identical results for estimating industrywide $R_j$. We observe that, unlike in the case of estimating totals, the methods are somewhat more competitive to each other except maybe for the jackknife method. One main reason for this seems to be that the point estimators for the ratios are much more stable across areas and PSUs and the correlations among the industries diminish, so that the effect of the sample design on their variance estimation is closer to that of stratified simple random sampling. Still, the jackknife estimator $v_{\text{Jack},R}$ is significantly positively biased in the case of estimating overall ratios for all three populations. The BHS estimator $v_{\text{BHS},R}$ is positively biased for the first population, and negatively biased for the third population, but is quite close to the "true" $V$ for the second population. The $v_R^*$ has a substantial positive bias for the first two populations, while the $\widetilde{v}_R$ works very well in contrast to its performance in the case of estimating totals. The general conclusion is that these estimators are not stable for all these cases. On the other hand, the proposed estimator $v_{\text{new},R}$ continues to perform well for all the populations.

The summary above is confirmed by the results of the empirical coverage probabilities and lengths of the corresponding confidence intervals given in Table 8. Recall that the standard error of the empirical coverage probabilities is about 0.0069. Therefore, the confidence interval using $v_{\text{BHS},R}$ is on target for the second population, but not so much for the other two populations. The confidence intervals using $v_{\text{Jack},R}$ are clearly too wide. Meanwhile, the confidence intervals using $v_{\text{new},R}$ have reasonably good coverage probabilities and lengths for all three populations. It is worth noting that the empirical coverage probabilities of the new method are exactly the same for both cases of estimating overall totals and ratios for all

three populations. It is a coincidence: the corresponding empirical coverage probabilities based on the first 500 runs are not all the same.

In the simulation study, we have also considered using the expansion in (20) with $\widehat{D}$ for the BHS and jackknife estimators. One significant observed change is that for the third population the jackknife estimator now underestimates the variance of the overall ratio estimator, behaving very closely to the BHS estimator which had little change from the standard version.

All the simulation experiments were conducted in Splus. The Splus code is available upon request.

## 6. Concluding remarks

Complex sample designs such as the one considered in this paper can often be thought of as *either* multi-stage or multi-phase. From the former point of view, we select different *types* of units at each stage. From the latter point of view, the units are always the ultimate sample units, and at each phase we gather more and more information, about narrower and narrower subsets of the population. Särndal et al. (1992, Chapter 9) regard multi-phase as subsuming multi-stage, and offer a general formula for two-phase variance estimation. Not surprisingly, a good deal of the subsequent literature has tended to emphasize the two-phase aspect, with a major question being the existence of simply applied replication variance estimation methods in this more general context.

Two referees point out that the relatively poor showing of the jackknife and BHS estimators is perhaps not surprising, when a key element of the design is that sampling of the second phase units is done according to a different classification than that dictating the first-phase selection, and there is a lack of the usual independence and invariance. In particular, the estimator of total we have considered here is a version of KS's "double expansion estimator" (DEE), and a key point is that KS find no satisfactory version of a jackknife estimator. They show through a simple example that standard versions of the jackknife cannot in general be consistent for the variance of the DEE estimator, as we have indeed found in our simulations.

Kim et al. (2000) offer a variant of the jackknife which is consistent for the DEE, in the particular case where the second phase of sampling is simple random sampling within second-phase strata. Binder et al. (2000) develop a Taylor series-based variance estimator, also limited to this case. It is at this point unclear whether their approaches extend to within stratum probability proportional to size sampling at the second phase, and how exactly this would be done. This looks like a promising avenue for further research. A referee has suggested that perhaps their estimators as they stand would be satisfactory, provided the one psu per stratum difficulty can be handled, and this is also worthy of exploration.

The hybrid method proposed in this paper makes use of both design-based and model-based strategies with the aim to solve difficult estimation problems arising from some complex, cross-stratified, random weighted two-phase sampling schemes, such as the NCS. We have provided technical justifications and some numerical support for this new approach. Judging from both the theoretical and empirical results, we see that the method appears to overcome the difficulties associated with this survey design. The new variance

estimator appears to be versatile, likely to be of use in practice. Finally, the ideas behind the construction of the hybrid estimator may be applied more generally to other settings of complex surveys.

## Acknowledgements

We would like to thank the editor and referees for their insightful and constructive comments and suggestions that led to significant improvements of this paper. In particular, to the referees we owe serious consideration of the two-phase aspect of the sample design considered, and the relevant literature. Wang's research was supported in part by the Bureau of Labor Statistics. Additional support was given by the National Cancer Institute (CA 57030), and by the Texas A&M Center for Environmental and Rural Health. The views in this paper are solely the authors' and do not reflect Bureau of Labor Statistics policy.

## References

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M., Jocelyn, W., 2000. Variance estimation for two-phase stratified sampling. Canad. J. Statist. 28, 751–764.

Cochran, W.G., 1977. Sampling Techniques, third ed. Wiley, New York.

Ernst, L.R., 1989. Weighting issues for longitudinal household and family estimates. In: Kaspszyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), Panel Surveys. Wiley, New York, pp. 139–159.

Folsom, R.E., Potter, F.J. Williams, S.R. 1987. Notes on a composite size measure for self-weighting samples in multiple domains. Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 792–796.

Kim, J.K., Navarro, A., Fuller, W., 2000. Variance estimation for 2000 census coverage estimates. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 515–520

Kott, P.S., 1990. Variance estimation when a first phase area sample is restratified. Survey Methodol. 16, 99–103.

Kott, P.S., Stukel, D.M., 1997. Can the jackknife be used with a two-phase sample?. Survey Methodol. 23, 81–89.

McCarthy, P.J., 1969. Pseudo-replication: Half-samples. Internat. Statist. Rev. 37, 239–264.

Quenouille, M.H., 1949. Approximate tests of correlation in time-series. J. Roy. Statist. Soc. Ser. B 11, 68–84.

Särndal, C.-E., Swensson, B., Wretman, J.H., 1992. Model Assisted Survey Sampling, Springer, New York.

Tukey, J., 1958. Bias and confidence in not quite large samples. Ann. Math. Statist. 29, 614.

Valliant, R., Dorfman, A.H., Royall, R.M., 2000. Finite Population Sampling and Inference: A Prediction Approach, Wiley, New York.

Wolter, K.M., 1985. Introduction to Variance Estimation, Springer, New York.